

AMS Seminar Series, August 21, 2018

Intelligent Databases and Machine Learning Forecast of Solar Flares

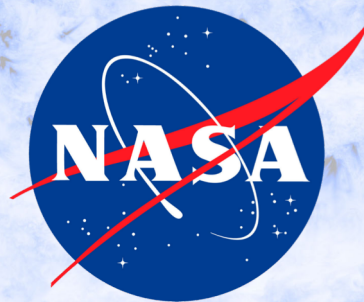
Viacheslav M Sadykov

Department of Physics, NJIT

Center for Computational Heliophysics, NJIT

Bay Area Environmental Research Institute

NASA Ames Research Center



Bay Area
Environmental

Research
Institute

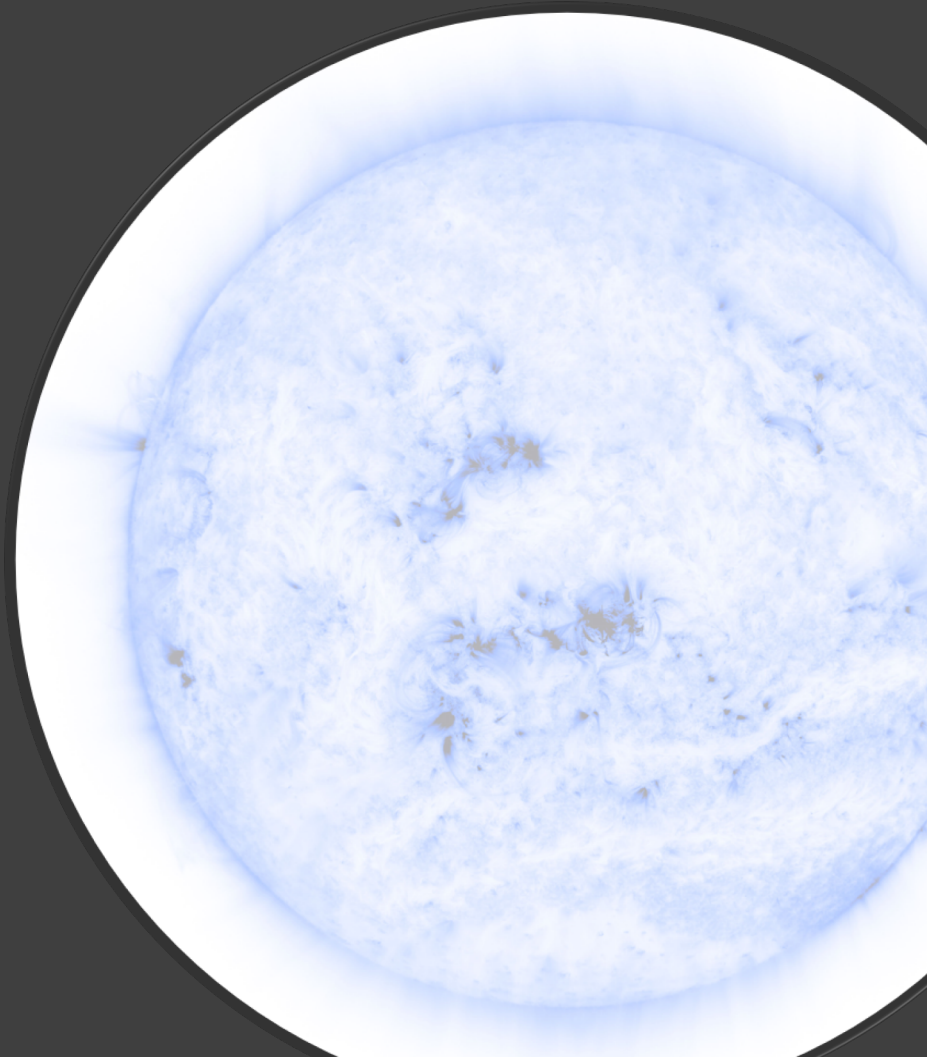
Outline

Introduction.
Solar flares and Space Weather.

Part I. Heliportal.
A new home for Interactive Multi-Instrument Database
of Solar Flares.

Part II. Predicting Solar Flares.
Can machine learning potentially provide a reliable
forecast?

Conclusions.
Large undiscovered scientific data volumes.



Solar Flares



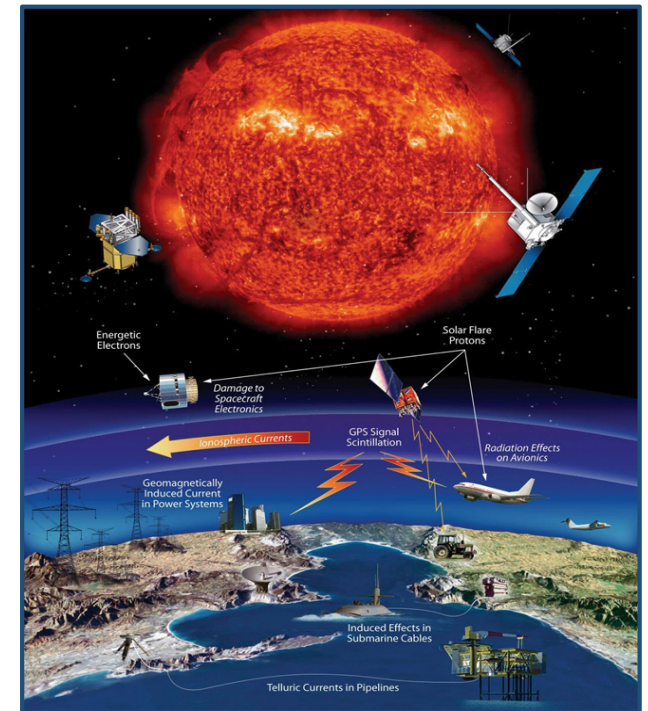
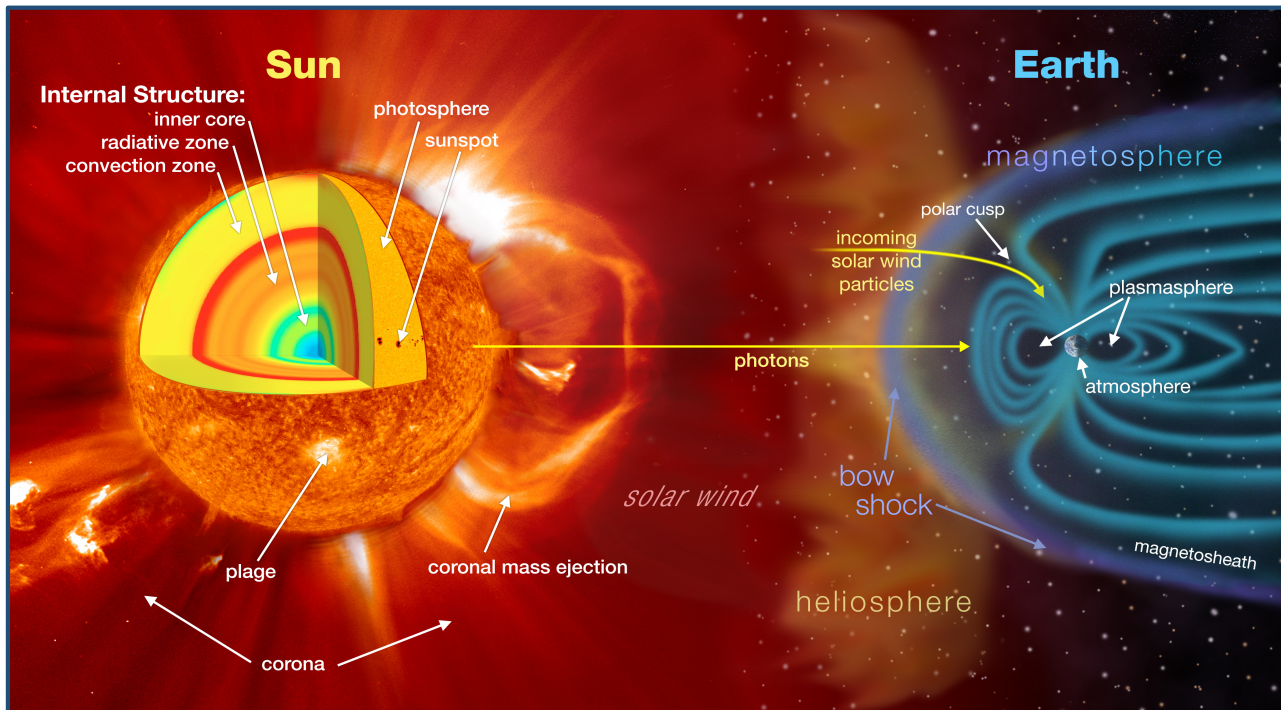
- Solar flares are the strongest transient energy release phenomena in the Heliosphere
- Tremendous amount of energy (up to 10^{32} erg) is released during one event in a timescale of tens of minutes
- Energy released during one solar flare is enough to cover the world energy consumption for 10 thousand years!



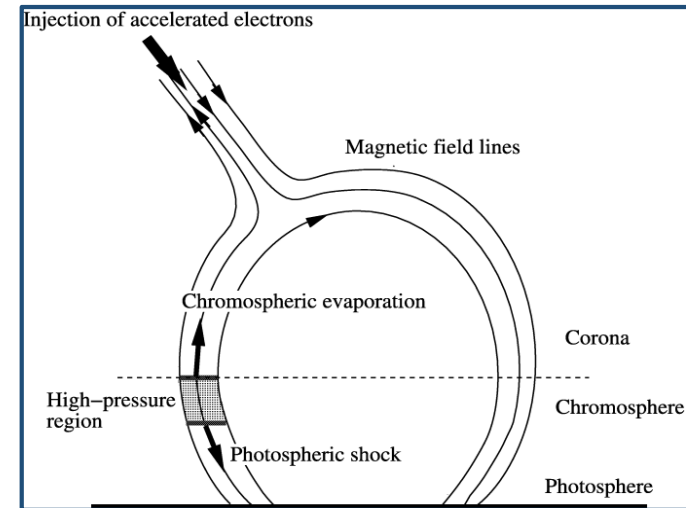
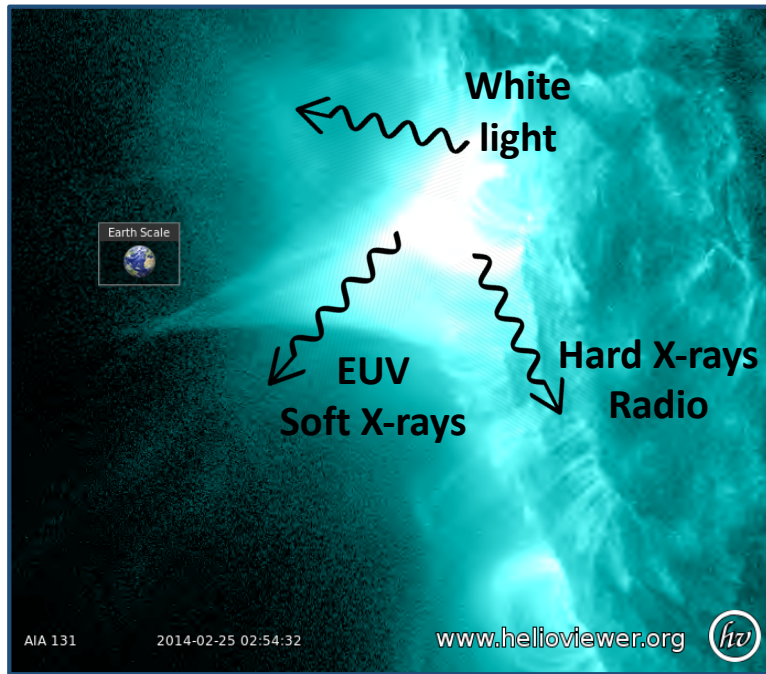
Impact on Space Weather

Solar flares represent the most prominent manifestation of the Sun's magnetic activity. They are often accompanied by high-speed coronal mass ejections and high-energy particles that greatly impact Earth's space environment and space weather, technological and biological systems:

- Destroy satellites equipment
- Affect radio communications and GPS navigation
- Disrupt power grids by return currents
- Provide potential danger for space exploration



Multiwavelength nature of flares



Standard model of solar flares includes physical processes reflected in different wavelengths / observation types:

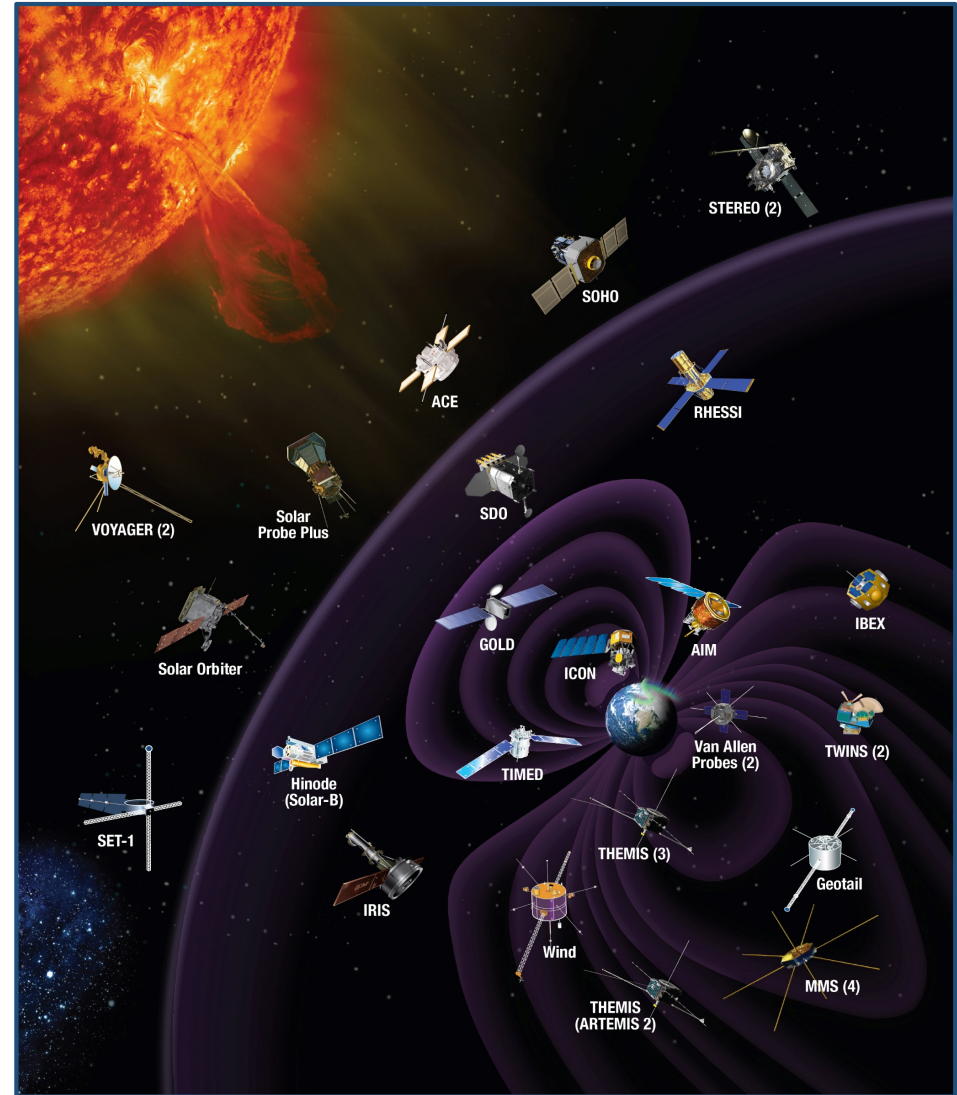
- **Magnetic reconnection in the coronal current sheet** (current sheet structure observed in **EUV**; termination shocks reported in **radio**)
- **Propagation of accelerated particles down into the chromosphere** (gyro synchrotron **radio** emission; **hard X-ray** bremsstrahlung emission; sometimes photospheric **white-light** emission)
- **Expansion of the heated chromospheric plasma into the coronal loops** (chromospheric evaporation; visible in **EUV images and spectra**, and recognized in **soft X-ray** emission behavior)
- **Eruptions of the plasma into the interplanetary space** (coronal mass ejections; visible in **EUV** images and coronagraph observations)

Observing Solar Flares

Solar flares cover the whole range of electromagnetic radiation spectrum, from radio- to gamma-rays, observed by many NASA space missions, including:

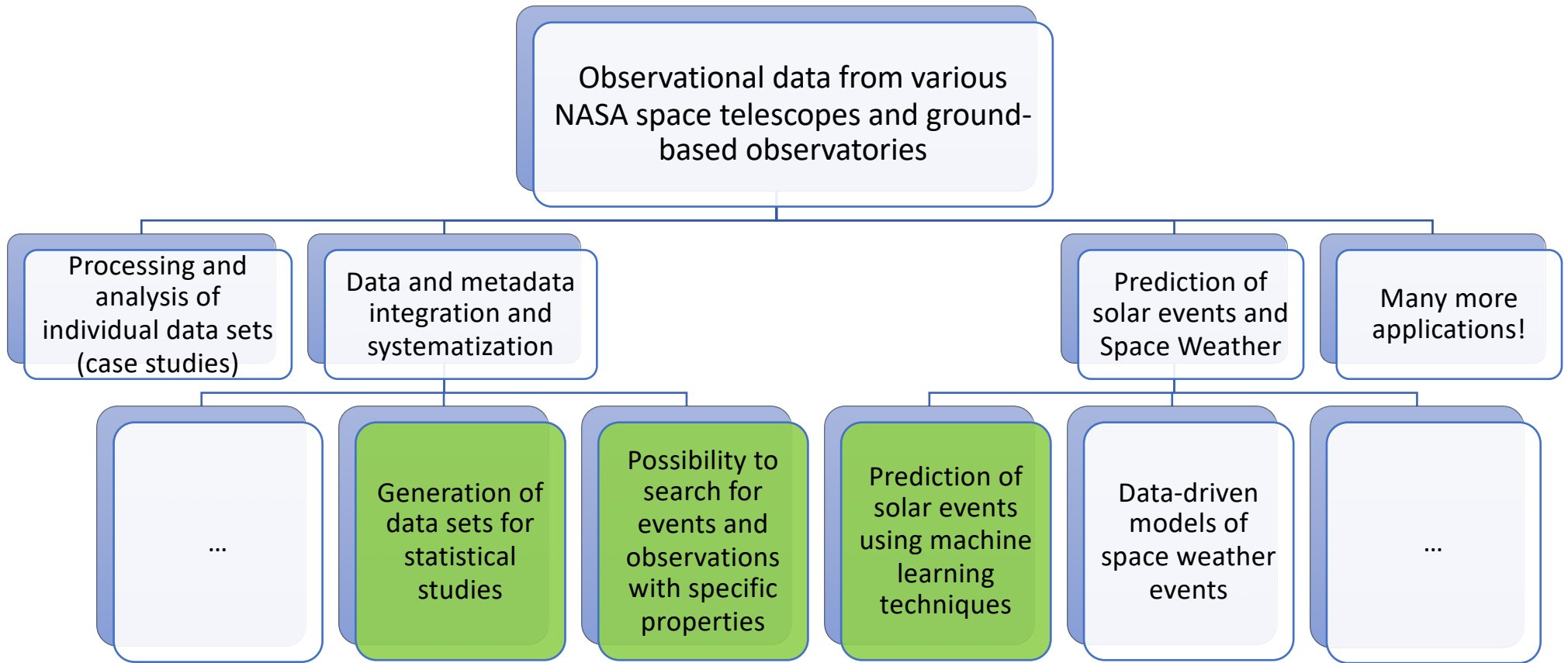
- Solar Dynamics Observatory (SDO)
- Geostationary Operational Environmental Satellite (GOES)
- Solar and Heliospheric Observatory (SoHO)
- Solar Terrestrial Relations Observatory (STEREO)
- Reuven Ramaty High Energy Solar Spectroscopic Imager (RHESSI)
- Interface Region Imaging Spectrograph (IRIS) ...

For a complete understanding of the flares it is necessary to perform a combined multi-wavelength analysis and classify large amounts of scientific data produced by space-based and ground-based observatories. Such classification and analysis will allow us to get clearer physical picture the observed phenomena, from their onset to space weather impacts.



Heliophysics Systems Observatory. Credits: <https://www.nasa.gov/>

Advantage of observational data

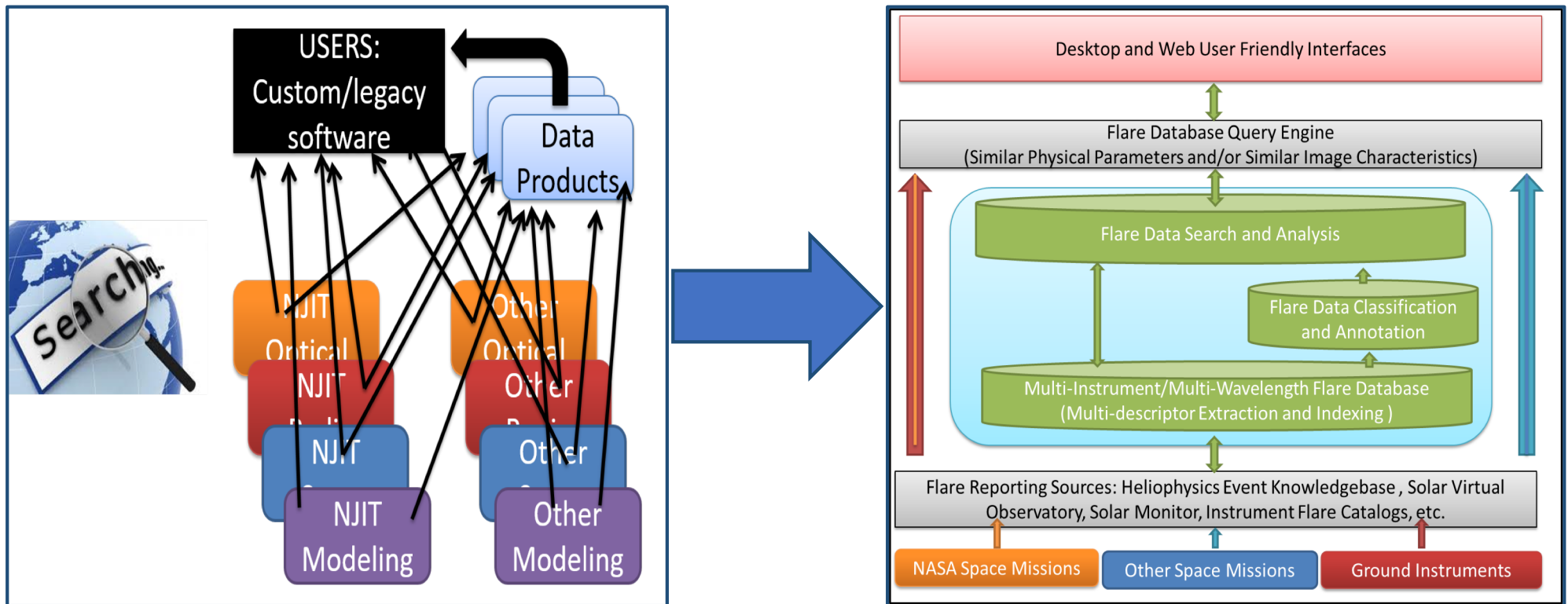




Part I. Heliportal

Motivation

- Many problems in solar-terrestrial physics require analyses to be performed using data from a particular set of instruments, and/or for a sample of flares with particular characteristics.
- Currently flare records made by different instruments are stored separately from each other. Matching records from different flare lists and search for the flare events with specific physical characteristics are complicated tasks...



Announcement of Heliportal (IMDSF)

An Interactive Multi-Instrument Database of Solar Flares (IMDSF, <https://solarflare.njit.edu/>) was developed for efficient data search, integration of different flare lists, and representation of observational data. IMDSF is fully functional and allows users to search for uniquely identified flare events based on physical characteristics (descriptors) and availability of observations of a particular set of instruments.

As part of a larger effort to provide a data portal for collaborative heliophysics research, NAS Division staff integrated and added security checks to the Interactive Multi-Instrument Database of Solar Flares in a new heliophysics portal, hosted on the NAS website: <http://heliportal.nas.nasa.gov>.

The IMDSF is a collaborative project among the NASA Advanced Supercomputing (NAS) Division, NASA Ames Heliophysics Modeling and Simulation team, and the New Jersey Institute of Technology's Department of Physics, Department of Computer Science and the Center for Computational Heliophysics.

Heliportal Milestones

June 2015: Project receives NASA support

Design and Implementation of a Multi-Instrument Database of Solar Flares (NASA NNX15AN48G, PI Gelu Nita, Co-Is: Alexander Kosovichev and Vincent Oria)

April 2016: Space Weather Workshop

First live demonstration of the Database of Solar Flares (E-poster)

July 2017: The IMDSF paper is published in ApJS

Sadykov V.M., Kosovichev A.G., Oria V., and Nita G.M. “*An Interactive Multi-instrument Database of Solar Flares*”. 2017, The Astrophysical Journal Supplement Series, Volume 231, Issue 1, article id. 6.

February 2018: The Heliportal is launched

The project is open for public, <https://heliportal.nas.nasa.gov>

First presentation of the Database of Solar Flares. Workshop supported by 2015 Faculty Seed Grant from NJIT, PI Alexander Kosovichev

January 2016: NJIT-NASA Workshop on Computational Heliophysics


The Interactive Multi-Instrument Database of Solar Flares (IMDSF, <https://solarflare.njit.edu/>) is released. The launch is announced in Solarnews.

February 2017: The IMDSF launch is announced

August 2017: IMDSF transfer to Heliportal is started

Interactive Multi-Instrument Database of Solar Flares

(IMIDSF, <https://heliportal.nas.nasa.gov/>, <https://solarflare.njit.edu/>)



Interactive Multi-Instrument Database of Solar Flares

(Click here to explore further.)

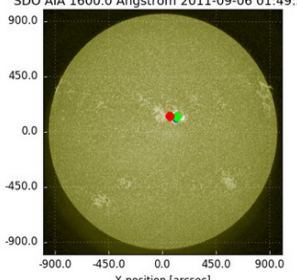
NAS ☐ NASA Ames
Questions / Feedback

[About](#) | [Query Page](#) | [Data Sources](#) | [Data Products](#) | [Contacts](#) | [Help](#)

Interactive Solar Flare Database

The Interactive Multi-Instrument Database of Solar Flares integrates records from various flare lists and catalogs, and allows the user to select the flare events based on their physical characteristics, ...

GOES Flares RHESSI Flares HEK Flares
SDO AIA 1600.0 Angstrom 2011-09-06 01:49:53



X-position [arcsec]

Project Description

The fundamental motivation of the project is that the scientific output of solar research can be greatly enhanced by better exploitation of the existing solar/heliosphere space-data products jointly with ground-based observations.

Our primary focus is on developing a specific innovative methodology based on recent advances in "big data" intelligent databases applied to the growing amount of high-spatial and multi-wavelength resolution, high-cadence data from NASA's missions and supporting ground-based observatories.

Our flare database is not simply a manually searchable time-based catalog of events or list of web links pointing to data. It is a preprocessed metadata repository enabling fast search and automatic identification of all recorded flares sharing a specifiable set of characteristics, features, and parameters.

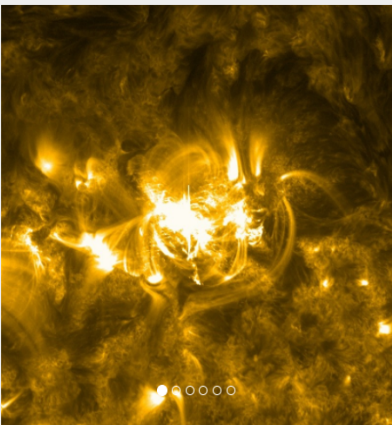
The result is a new and unique database of solar flares and data search and classification tools for the Heliophysics community, enabling multi-instrument/multi-wavelength investigations of flare physics and supporting further development of flare-prediction methodologies.

[Launch Solar Flare Query Page](#)

[ABOUT](#) | [DATA SOURCES](#) | [QUERY](#) | [DATA PRODUCTS](#) | [CONTACTS](#)

Interactive Multi-Instrument Database of Solar Flares

Click to explore



Team: Viacheslav Sadykov, Rishabh Gupta, Dr. Alexander Kosovichev, Dr. Vincent Oria, Dr. Gelu Nita

Project Description

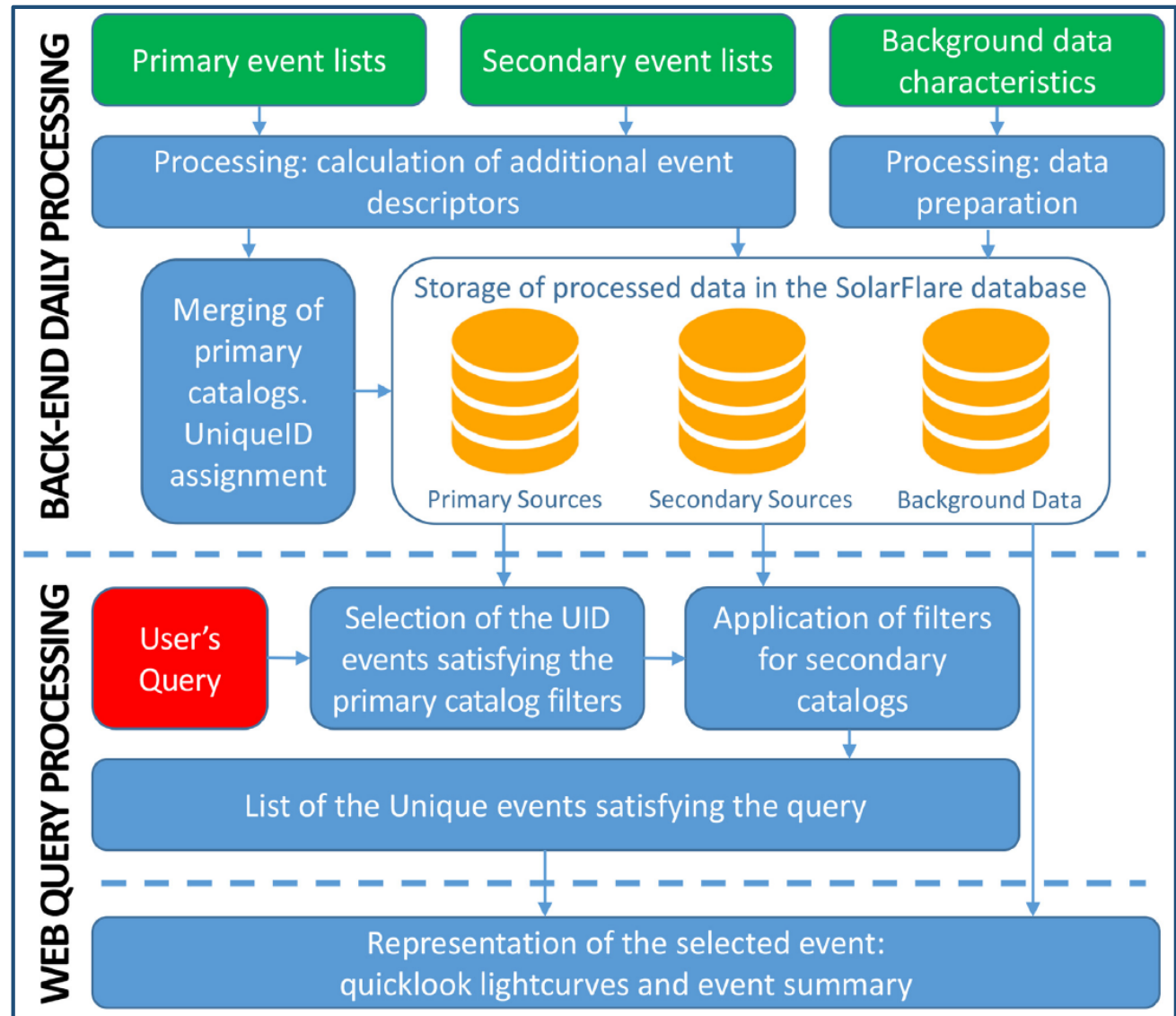
The fundamental motivation of the project is given by the idea that the scientific outcome of solar research can be greatly enhanced by a better exploitation of the existing solar/heliosphere space data products jointly with ground-based observations.

Our primary focus is on developing a specific innovative methodology based on the recent advances in "big data" intelligent databases and on the tremendously growing amount of high-spatial and multi-wavelength resolution, high-cadence data from NASA's missions and supporting ground-based observatories.

IMIDSF Structure

Heliportal consists of several functional elements:

1. **Back-End MySQL database** containing primary and secondary event lists and background data sources
2. **Back-End daily-update system (PHP and Python scripts)** for data upload, processing and enrichment
3. **Front-End web application** with the user query form and presentation of the query results and event summary



Heliportal Data Sources

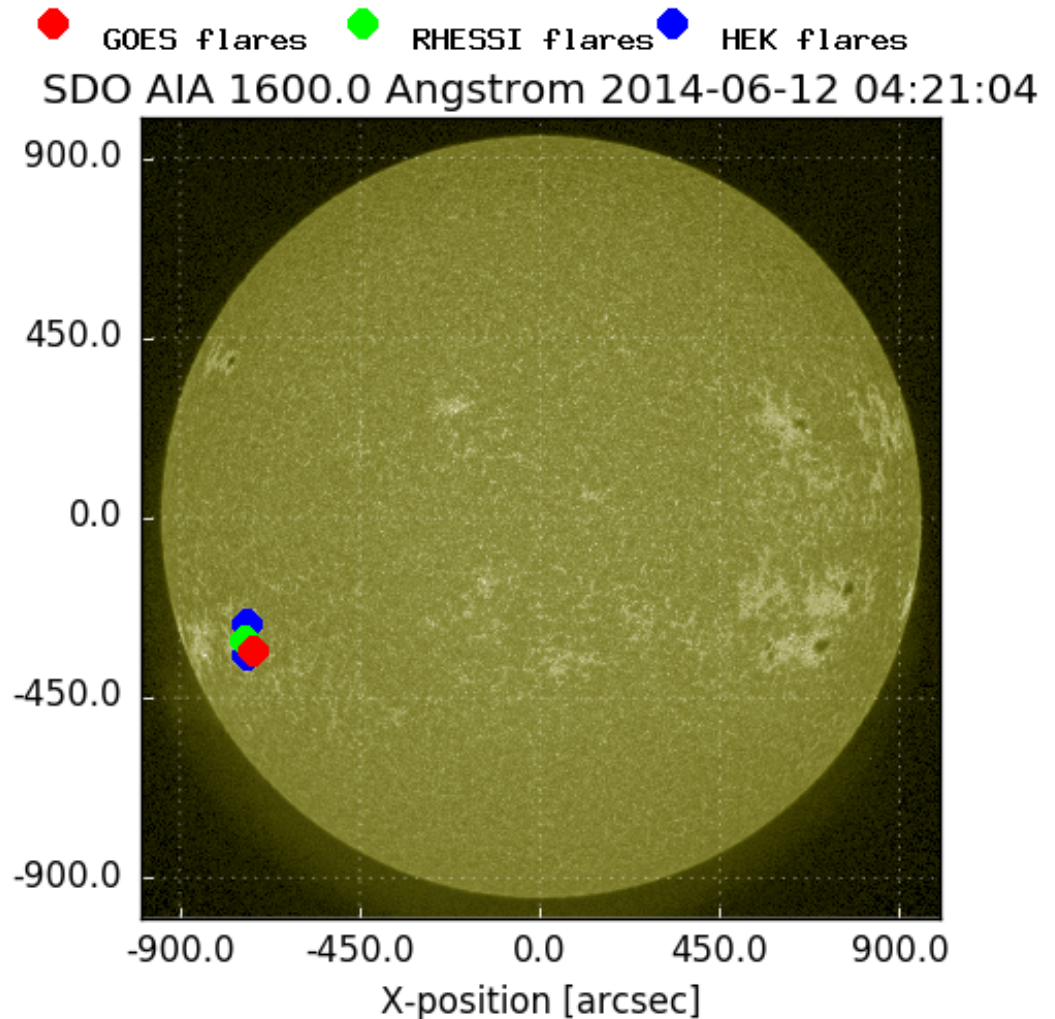
Primary Event Lists		
GOES flare list	2002 Jan – current time	ftp://ftp.swpc.noaa.gov/pub/warehouse/
RHESSI flare list	2002 Feb – current time	http://hesperia.gsfc.nasa.gov/hessidata/dbase/
HEK flare list	2010 Feb – current time	https://www.lmsal.com/isolsearch
Secondary Event Lists		
IRIS observing logs	2013 Jul – current time	http://iris.lmsal.com/search/
Hinode flare catalog	2006 Nov – 2016 Jul	http://st4a.stelab.nagoya-u.ac.jp/hinode_flare/
Fermi GBM flare catalog	2008 Nov – current time	https://hesperia.gsfc.nasa.gov/fermi/gbm/qlook/
Nobeyama coverage check	2010 Jan – current time	ftp://solar-pub.nao.ac.jp/pub/nsro/norp/xdr/
OVSA flare catalog	2002 Jan – 2003 Dec	http://www.ovsa.njit.edu/data/
EOVSA flare catalog	2017 Jan – current time	http://www.ovsa.njit.edu/wiki/index.php/Expanded_Owens_Valley_Solar_Array
CACTus CME catalog	2002 Jan – current time	http://sidc.oma.be/cactus/
Filament eruption catalog	2010 Apr – 2014 Oct	http://aia.cfa.harvard.edu/filament/
Konus-WIND flare catalog	2002 Jan – 2016 Jul	http://www.ioffe.ru/LEA/Solar/index.html
Background Data Characteristics		
GOES X-ray light curves (and T&EM)	2002 Jan – current time	https://umbra.nascom.nasa.gov/goes/fits/
SDO/EVE ESP light curves	2010 Feb – current time	http://lasp.colorado.edu/eve/data_access/
Nobeyama Polarimeter data	2010 Jan – current time	ftp://solar-pub.nao.ac.jp/pub/nsro/norp/xdr/

Heliportal Data Enrichment and Processing

Our daily-based processing includes:

- Identification and assignment of unique identifiers (UniqueIDs) for flares from GOES, RHESSI, and HEK flare catalogs
- Determination of missing coordinate information from the position of parental active regions
- Calculation of some data products (example: temperatures and emission measures based on GOES observations in 0.5-4 Å and 1-8 Å channels)

Counterparts for gev_20140612_041400
M2.0 class flare from GOES, RHESSI
and HEK flare catalogs



Examples of Application

Statistical Study of Chromospheric Evaporation in Solar Flares

- To connect energy fluxes deposited in solar flares and the properties of the responding solar plasma and compare results with the RHD chromospheric evaporation simulations, **the dataset of flares simultaneously observed by IRIS** (here in the fast-scanning regime) **and RHESSI is required**

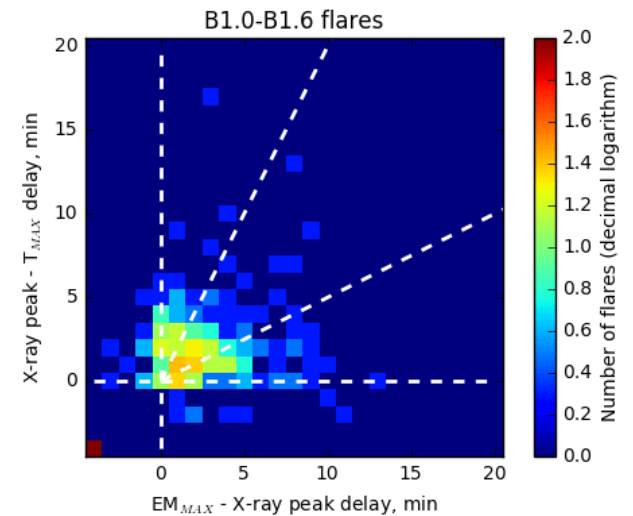
Statistical study of Soft X-ray Emission Properties and Timescales from GOES observations

- The results of the application of TEBBS algorithm (T and EM calculations) for GOES flares detected from 2002 until today **are available as a data product at Heliportal** (<https://heliportal.nas.nasa.gov/>). **IMDSF allows us to integrate GOES and RHESSI flares** and catch the difference between the flares with different timescale relations.

Forecasting of Solar Flares using Machine-Learning Methods

- Heliportal allows the users to request **the statistics of flares for each AR** in one click. Integration of the AR magnetic field descriptors (SHARP parameters and PIL parameters) with flare events is planned. It is also possible to request **not only the GOES class but other physical characteristics of solar flares**, and, in principle, work on the prediction of these characteristics.

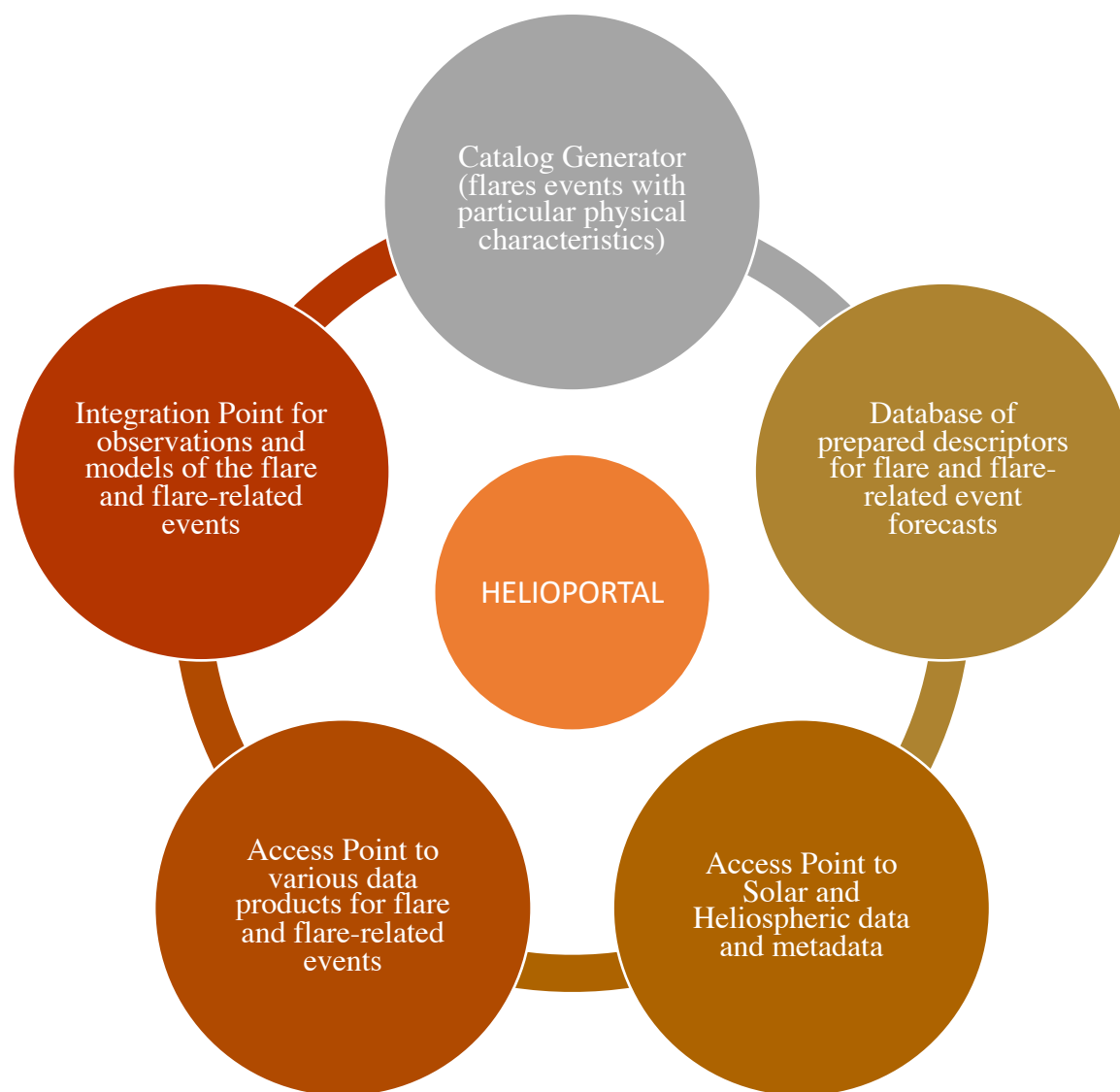
SOLID	GOES class	RHESSI energy range, keV	IRIS mode	IRIS cadence, s	Hinode coverage
SOL2014-02-13T01:32:00	M1.8	6-12	Coarse raster, 8 slits	41.7	XRT, EIS
SOL2014-02-13T02:41:00	M1.0	3-6	Coarse raster, 8 slits	41.7	XRT
SOL2014-03-29T17:35:00	X1.0	100-300	Coarse raster, 8 slits	71.9	SOT FG, XRT, EIS
SOL2014-06-12T18:03:00	M1.3	25-50	Coarse raster, 8 slits	21.3	XRT
SOL2014-06-13T00:30:00	C8.5	12-25	Coarse raster, 8 slits	21.3	-
SOL2015-03-11T11:21:00	C5.8	12-25	Coarse raster, 8 slits	75.0	SOT SP
SOL2015-11-04T13:31:00	M3.7	50-100	Coarse raster, 16 slits	49.5	-



Future Plans

Our long-term goal is to expand the functionality of the Heliportal. We are currently developing the Intelligent Database of Solar Events and Active Regions (IDSEAR). We plan to include the following into new database:

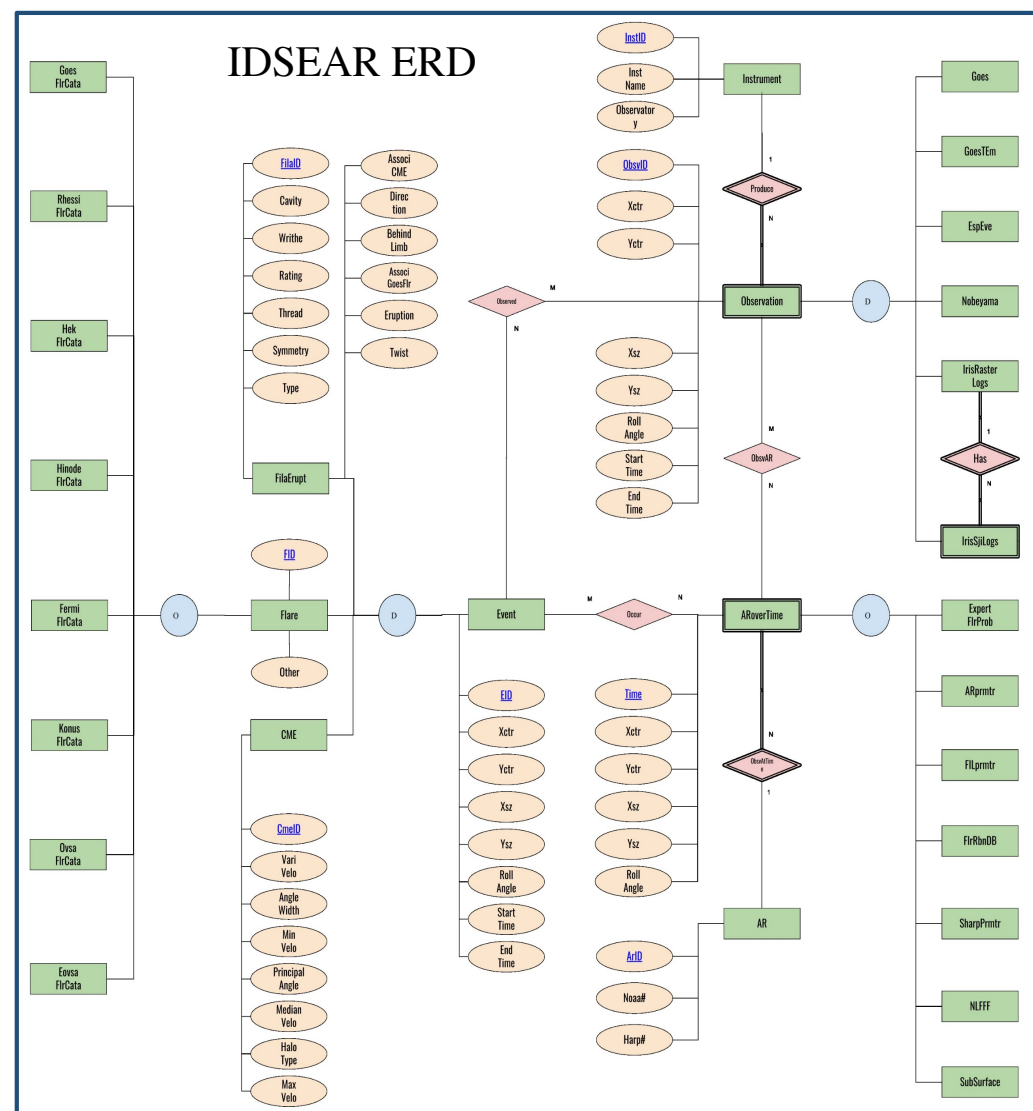
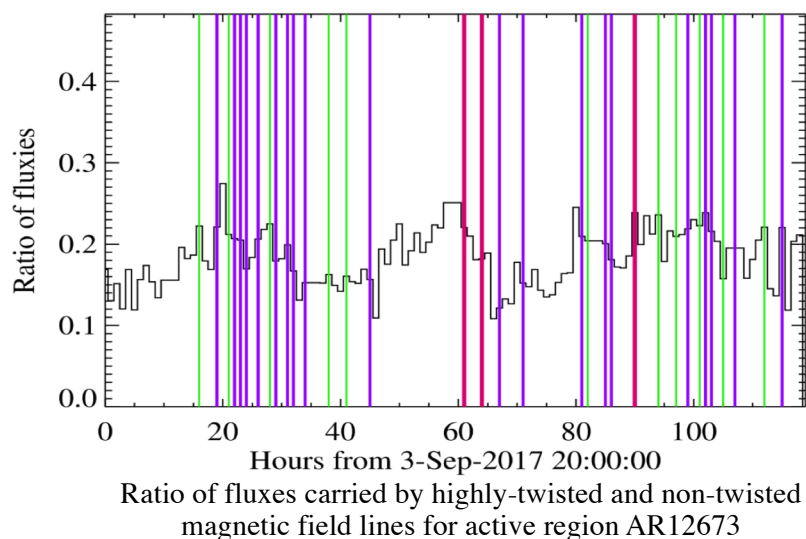
- Increase number of flare and flare-related event sources and observational characteristics/logs
- Include Active Region characteristics (PIL characteristics, SHARP parameters) and integrate them with solar events
- Integrate solar events with existing models; provide initial conditions for models
- Provide multi-level access to the database (possibility to work with both the products of integration and catalogs before integration; started)
- Provide various data products (subsurface flow maps, NLFFF extrapolations for flaring ARs etc)
- Develop IDL and Python packages to access the database, requests catalogs and data products
- Integrate with other resources



IDSEAR in progress

At the current stage, we have completed the following steps towards the IDSEAR implementation:

1. Designed IDSEAR and performed accurate mapping of the designed schema to the relational MySQL database
2. Transferred data from IMDSF into IDSEAR
3. Integrated records from different flare catalogs and assigned unique flare identifiers (UniqueIDs) for the flare records
4. Loaded SHARP data from JSOC/Stanford and PIL descriptors as tables under “Active region over time” entity
5. Implemented and tested codes for various AR data products and descriptors (NLFFF cubes, subsurface flow maps and related descriptors)



IDSEAR queries

IDSEAR DB already operates at NJIT server. Here are some examples of future “typical” queries to the DB.

Example 1.

- Properties of the parental active regions (SHARP parameters) for strong flares (M and X class) 12 hours before the flare.
- About 45 seconds for whole database request

```
MySQL [sun_rsrch]> SELECT s.SharpID, s.USFLUX, s.AREA_ACR, s.R_VALUE,
-> g.GoesID, g.StartTime, g.EndTime, g.FlareClass
-> FROM SharpPrmtr AS s FORCE INDEX (SharpID, T_REC, DATE_OBS, NOAA_AR),
-> GoesFlrCata as g FORCE INDEX (GoesID, StartTime_2, StartTime, EndTime, GoesARNum)
-> WHERE g.StartTime BETWEEN '2010-01-01' AND '2019-01-01'
-> AND g.FlareClass BETWEEN 'M1.0' AND 'X9.9'
-> AND g.GoesARNum <> 0
-> AND s.NOAA_AR = g.GoesARNum
-> AND timestampdiff(MINUTE,g.StartTime,s.T_REC) > 720
-> AND timestampdiff(MINUTE,g.StartTime,s.T_REC) < 732;
```

SharpID	USFLUX	AREA_ACR	R_VALUE	GoesID	StartTime	EndTime	FlareClass
2918	1.088371e22	258.373688	4.08	12165	2010-05-05 17:13:00	2010-05-05 17:22:00	M1.2
19085	1.016187e22	279.260132	3.991	12222	2010-06-12 00:30:00	2010-06-12 01:02:00	M2.0
43005	1.093167e22	634.591919	3.466	12397	2010-08-07 17:55:00	2010-08-07 18:47:00	M1.0
75585	1.480548e22	602.007629	4.179	12636	2010-10-16 19:07:00	2010-10-16 19:15:00	M2.9
88167	2.090904e22	402.275299	3.758	12763	2010-11-04 23:30:00	2010-11-05 00:12:00	M1.6
88233	2.787108e22	524.238647	4.094	12771	2010-11-05 12:43:00	2010-11-05 14:06:00	M1.0
2563316	5.970982e22	1812.145752	4.789	25847	2017-09-08 07:40:00	2017-09-08 07:58:00	M8.1
2563353	5.927371e22	1741.391846	4.371	25853	2017-09-08 15:09:00	2017-09-08 16:04:00	M2.9
2563392	5.349583e22	1759.272827	4.49	25858	2017-09-08 23:33:00	2017-09-08 23:56:00	M2.1
2563416	4.196586e22	1767.531128	4.795	25862	2017-09-09 04:14:00	2017-09-09 04:43:00	M1.1
2563449	2.078946e22	1153.984497	4.64	25866	2017-09-09 10:50:00	2017-09-09 11:42:00	M3.7
2563505	0	116.621277	5.089	25871	2017-09-09 22:04:00	2017-09-10 00:41:00	M1.1
2584868	5.852978e21	167.674744	3.474	25953	2017-10-20 23:10:00	2017-10-20 23:37:00	M1.1

564 rows in set (44.76 sec)

```
MySQL [sun_rsrch]> SELECT s1.SharpID, s1.HARPNUM, s1.T_FRST1, s1.T_REC, s1.LON_FWT, s1.LAT_FWT, s1.USFLUX
-> FROM SharpPrmtr AS s1 FORCE INDEX (SharpID, HARPNUM, T_FRST1_2, T_FRST1, T_REC, LON_FWT_2, LON_FWT, LAT_FWT)
-> WHERE s1.T_FRST1 BETWEEN '2010-01-01' AND '2019-01-01'
-> AND s1.LON_FWT BETWEEN -45.0 AND 45.0
-> AND s1.LAT_FWT BETWEEN -45.0 AND 45.0
-> AND (s1.LON_FWT <> 0 OR s1.LAT_FWT <> 0)
-> AND s1.T_FRST1 = s1.T_REC;
```


SharpID	HARPNUM	T_FRST1	T_REC	LON_FWT	LAT_FWT	USFLUX
1248	2	2010-05-01 00:00:00	2010-05-01 00:00:00	-28.227911	15.40951	1.846205e21
1751	5	2010-05-01 00:00:00	2010-05-01 00:00:00	7.801724	-28.844969	2.790045e20
1985	6	2010-05-01 00:00:00	2010-05-01 00:00:00	-2.004806	-31.786865	6.642536e20
2484	8	2010-05-02 14:36:00	2010-05-02 14:36:00	-1.990009	41.479801	1.0241e20
2612947	7250	2018-04-18 15:36:00	2018-04-18 15:36:00	-4.497038	-14.862375	2.562211e19
2619274	7261	2018-05-22 20:36:00	2018-05-22 20:36:00	10.969915	5.345559	1.344429e19
2621458	7265	2018-05-27 16:00:00	2018-05-27 16:00:00	22.245598	-4.783037	2.747221e19
2621565	7266	2018-06-03 01:00:00	2018-06-03 01:00:00	44.270744	22.077038	2.809222e20
2621586	7267	2018-06-05 07:48:00	2018-06-05 07:48:00	-29.342995	14.169471	4.55595e19
2623271	7274	2018-06-17 21:12:00	2018-06-17 21:12:00	24.734566	7.777799	2.733343e19

1609 rows in set (39.45 sec)

Example 2.

- List of active regions (SHARPs) appeared on the Sun at user-defined time, longitude, and latitude ranges
- About 40 seconds for whole table request

API and UI are coming soon...

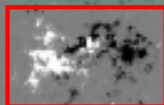
A blue-tinted image of the Sun's surface, showing a textured, granular appearance with several dark, irregular patches (sunspots) and bright, swirling regions (solar flares). The image is partially obscured by a white circular overlay on the right side.

Part II. Predicting Solar Flares

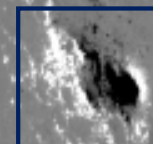
Where will an M-class flare happen?



0 M-class flares



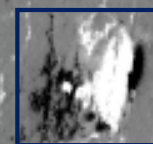
1 M-class flare



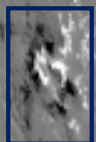
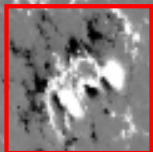
0 M-class flares

SWPC NOAA flare probability for this day: 70%

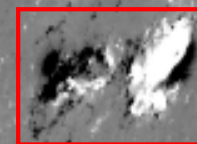
0 M-class flares



4 M-class flares



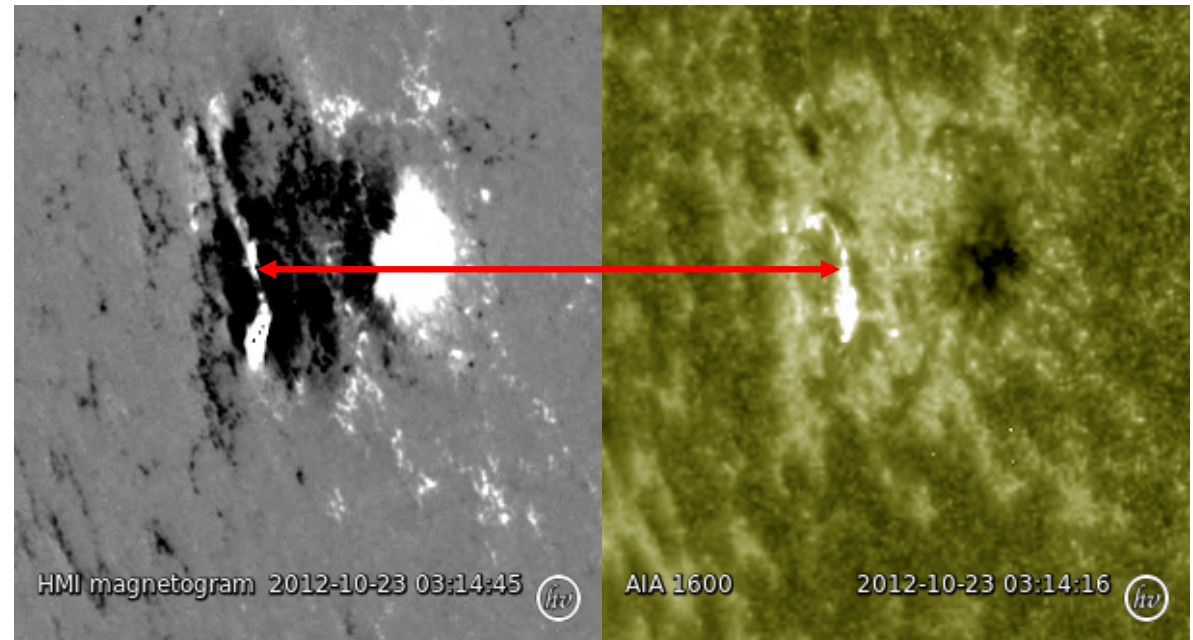
0 M-class flares



1 M-class flare

Why is it important to predict flares?

- The prediction of strong solar flares is one of the key questions of Solar Physics: solar flares are one of the primary drivers of space weather
- Besides many attempts, the operational predictions are still mainly done based on experts opinion and experience.
- Many observational studies and previous ML studies confirmed the important role of the properties of the magnetic field in active regions for prediction of solar flares
- Attempts to predict flare events help us to understand physics of solar flares and their triggering mechanisms

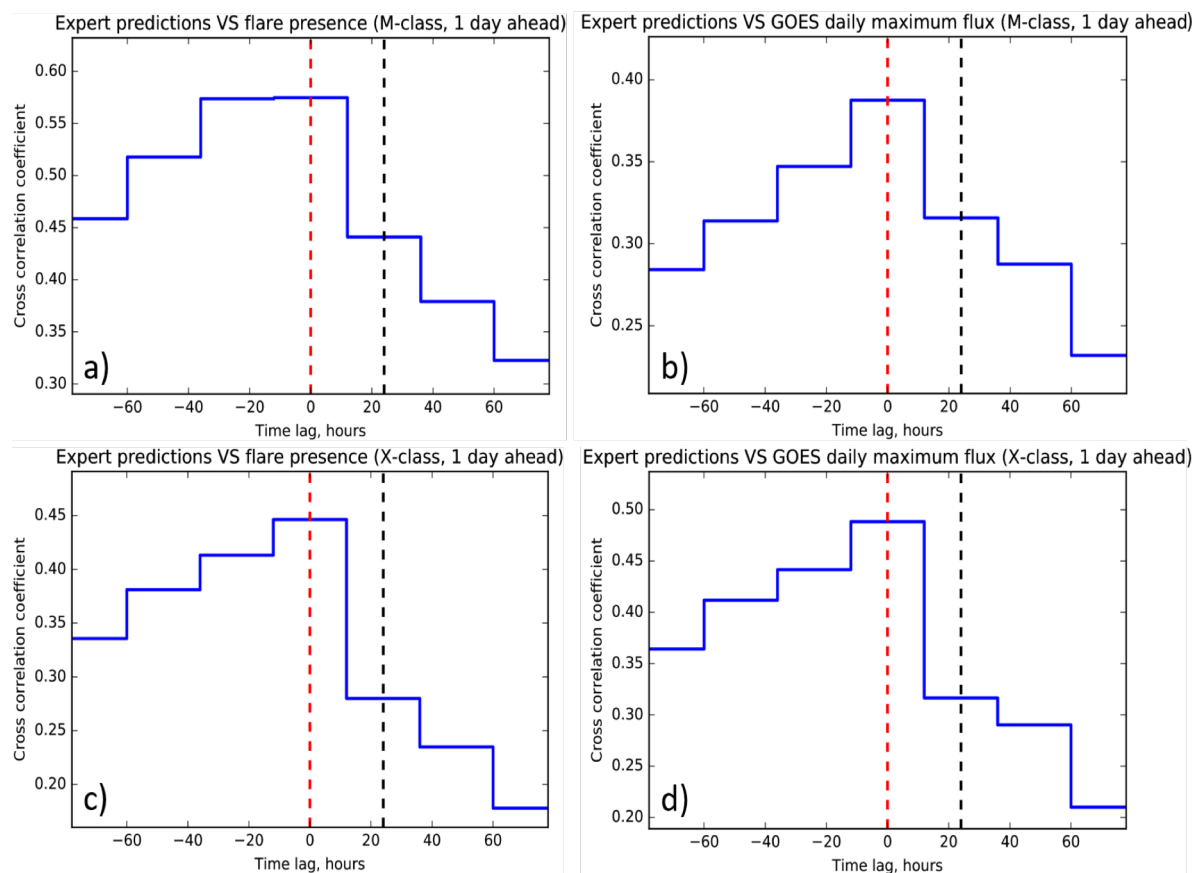


Credits: <https://helioviewer.org/>

Primary goal: utilize an advantage of high volumes of observational data to discover new flare-sensitive features, evaluate importance of particular types of observations, and potentially enhance operational forecasts

Properties of the SWPC NOAA Operational Forecasts

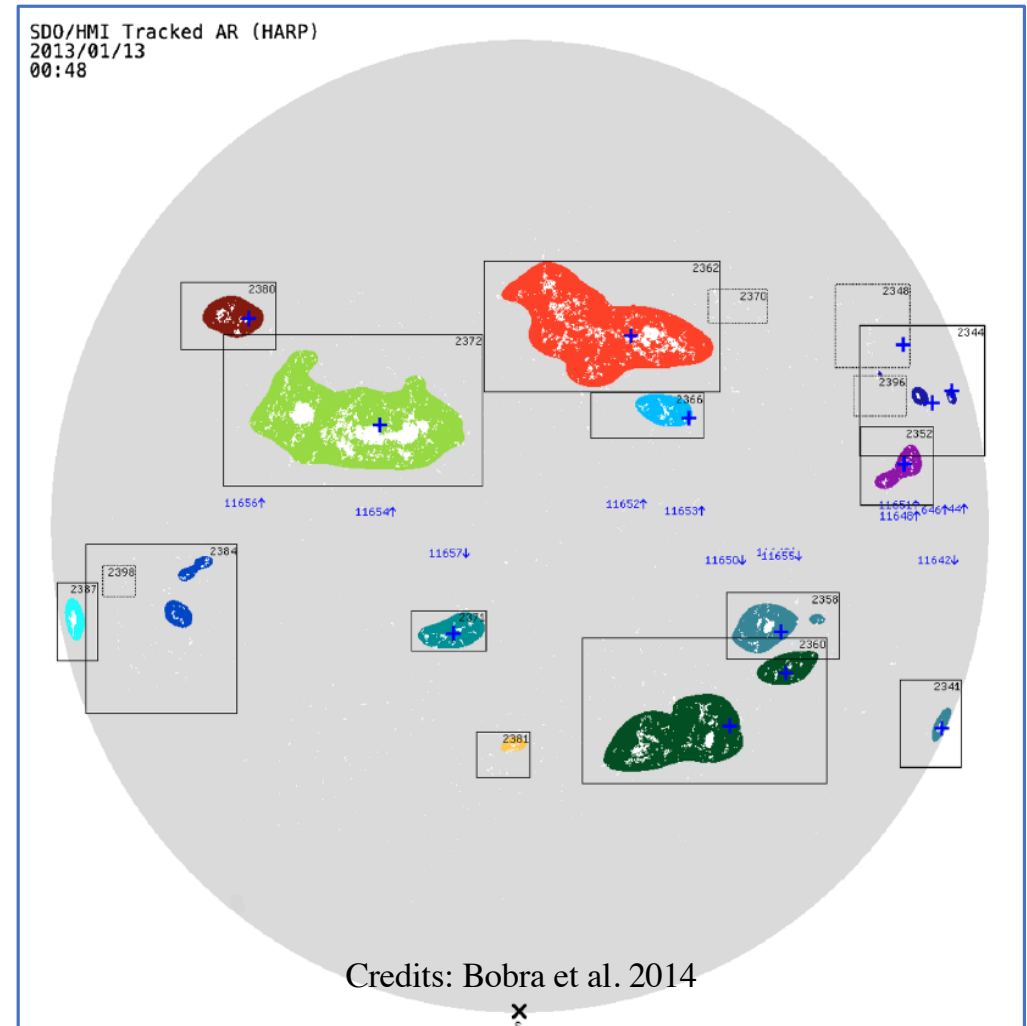
- The current daily operational forecasts at the SWPC are made by forecasters for each of the three upcoming days using a modified three-component Zurich class (McIntosh 1990) and magnetic class (Smith & Howard 1968) for each active region and historical look-up tables of flare probability as a function of active region class, flaring history, growth/decay of sunspots. The calculated probability is corrected by forecasters based on their experience.
- The flare prediction probabilities are correlated stronger with the current flare activity than with the next-day activity
- Nevertheless, the expert-based probabilities represent valuable information for the flare forecast.



Cross correlation coefficient of the expert probabilities of M-class and X-class flares and various SXR characteristics of the flare activity

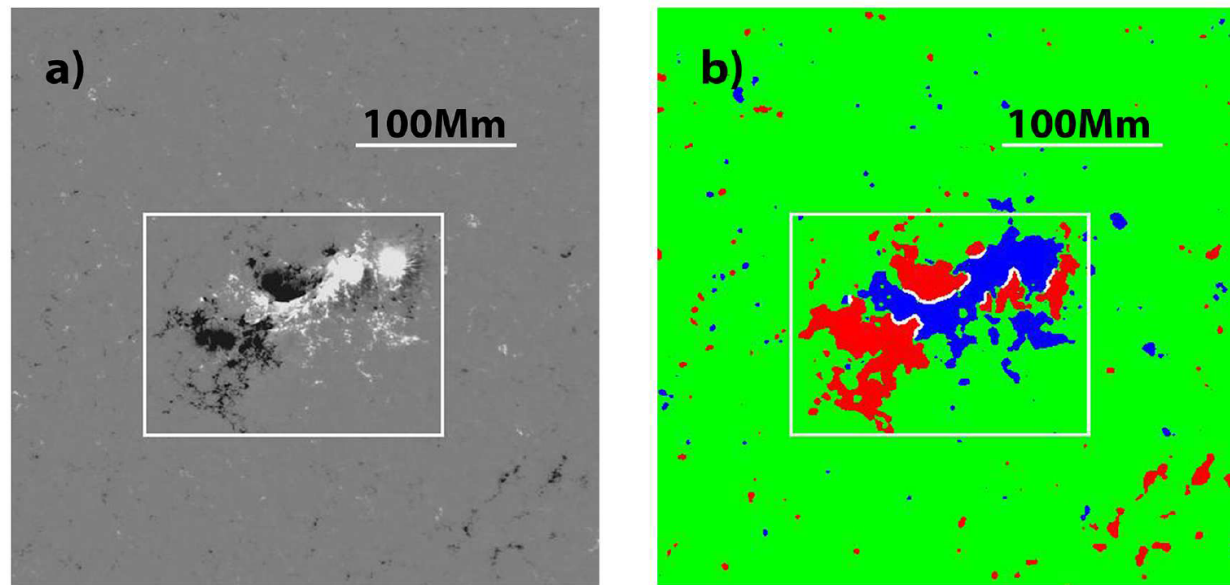
Magnetic field in active regions

- The magnetic field is the only reservoir able to store the typical energy released during the solar flares => one has to look at magnetic properties of parental active regions
- NASA's Helioseismic and Magnetic Imager onboard the Solar Dynamics Observatory (SDO/HMI) provides the routine coverage of the whole Sun photospheric line-of-sight and vector magnetic field data since 2010, resulted in more than 1PB generated data
- Examples of magnetic field descriptors in Active Regions:
 - Space weather HMI Active Region Patches (SHARPs, Bobra et al. 2014)
 - Properties of the magnetic field Polarity Inversion Line (PIL) in strong field regions (Sadykov and Kosovichev 2017)
 - Descriptors of extrapolated 3D magnetic field structure (free energy excess, ratio of fluxes in twisted/untwisted lines, ...)



PIL detection algorithm and extracted features

AR 11158 2011-02-16T20:00:00



- Previous statistical and case studies of solar flares demonstrated importance of the magnetic polarity inversion lines (PILs) in active regions for the flare initiation and development process
- We divide the line-of-sight active region magnetogram into regions with strong positive field (“positive” segments), strong negative field (“negative” segments), and weak field (“neutral” segments).
- For each AR, we remap the LOS magnetogram onto the heliographic coordinates, and solve the segmentation problem formulated as an optimization task (Chernyshov et al, 2011).
- The segmentation results are used to determine the PIL and corresponding characteristics

Schema for the binary flare forecast in active region

1. Construction of labeled data set

- Measure characteristics of the active region (SHARPs, PIL) at a certain time moment
- Determine if a strong flare happened in the active region within certain time (say, 24 hours) from the considered moment
- One has: vector of characteristics and its label (0 or 1)

2. Separation of data into train/validation/test data subsets

3. Feature selection on train data set (F-score, Gini importance, other)

4. Optimization of the classifier on train/validation data sets

- Different classifiers have different inner parameters which should be optimized
- Target: maximization of a certain metrics. Example: $TSS = \frac{TP}{TP+FN} - \frac{FP}{FP+TN}$

5. Performance of the classifier on test data set. Comparison of results.

Examples of previous results (TSS)

Sadykov and Kosovichev (2017)		PIL characteristics only	PIL + global characteristics	50% decreased threshold values
Prediction of $\geq M1.0$ flares		0.76 ± 0.03	0.74 ± 0.03	0.76 ± 0.03
Prediction of $\geq X1.0$ flares		0.84 ± 0.07	0.84 ± 0.07	0.85 ± 0.04
OTHER WORKS	Bobra & Couvidat (2015) (vector MF)	Nishizuka et al (2017) (vector MF, flare prehistory etc.)		Nishizuka et al (2018) (operational separation of DS)
$\geq M1.0$ flares	0.82 (SVM)	0.87 (SVM), 0.91 (kNN)		0.80 (DNN), 0.33 (SVM)
$\geq X1.0$ flares	-	0.88 (SVM), 0.91 (kNN)		-
EXPERT-BASED PREDICTIONS	NICT SWFC (from Nishizuka et al 2017)		Royal Observatory of Belgium (from Nishizuka et al 2017)	NOAA SWPC (from Crown 2012, Table 4)
$\geq M1.0$ flares	0.50		0.34	0.53
$\geq X1.0$ flares	0.21		-	0.49

For more accurate comparison with expert-based predictions (any operational forecast) one needs at least to unify dataset structures. We attempt to do it by obtaining daily descriptors and predicting next day flare activity (as done by SWPC).

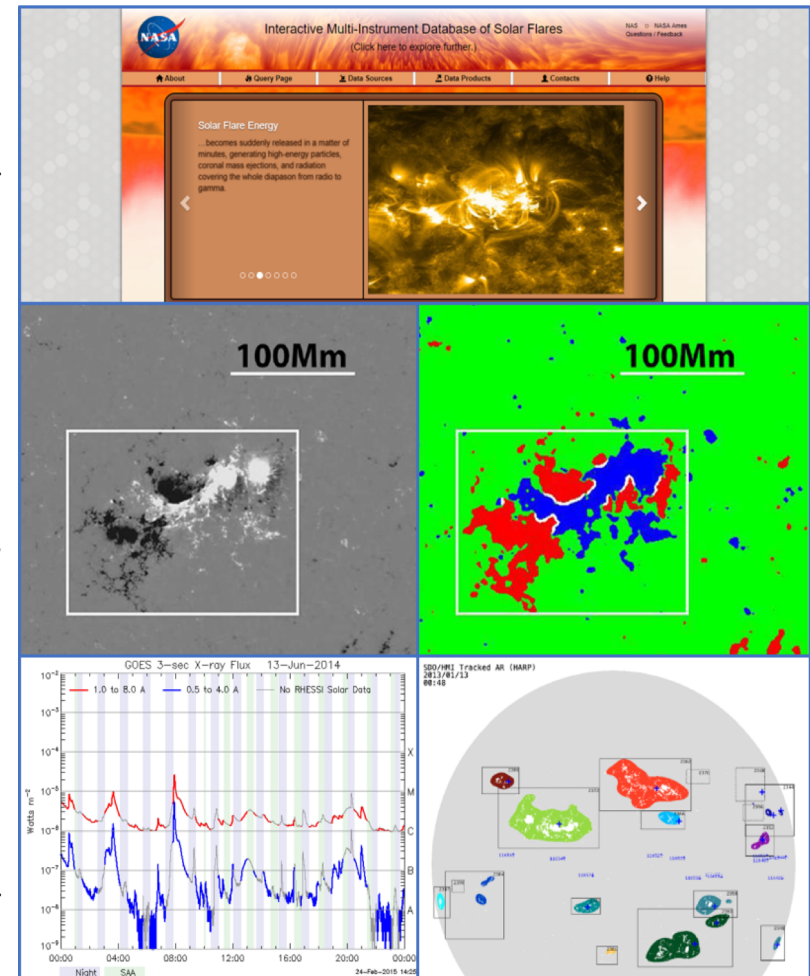
Primary goal: investigate the possibility of enhancement of the SWPC NOAA operational forecasts by employing machine-learning algorithms to combine expert predictions with magnetic field and soft X-ray flux characteristics

Our study: Data Sources and Descriptors

1. SWPC NOAA operational forecasts (probabilities) of M/X-class flares for the next day (<ftp://ftp.swpc.noaa.gov/pub/warehouse/>)
2. Statistics of M/X-class flares from <https://heliportal.nas.nasa.gov/>
3. SXR 1-8Å flux obtained by GOES/XRS
4. Polarity Inversion Line (PIL) characteristics obtained from SDO/HMI line-of-sight magnetic field data (Sadykov and Kosovichev, 2017)
5. Space Weather HMI Active Regions Patches for NOAA ARs (SHARPs, Bobra et al. 2014)

The data are obtained for May 01, 2010 – Dec 31, 2017 time period. For each day for the midnight time, we obtain the following features of the solar activity:

- Averaged and peak SXR fluxes during the 1-3 preceding days
- Total number of M-class and X-class flares during the 1-3 preceding days
- Daily mean and maximum values of the PIL characteristics (maxima over ARs are selected)
- Daily mean and maximum values of the SHARP characteristics (maxima over ARs are selected)



Credits: <https://heliportal.nas.nasa.gov/>, Sadykov and Kosovichev 2017, Bobra et al. 2014, RHESSI Browser (<http://sprg.ssl.berkeley.edu/~tohban/browser/>)

Feature Selection Algorithm

1. The labels for the data set are assigned: 1 if there is an M/X-class flare which happened on the next day, 0 otherwise. The days when the flares were only located close to the limb are ignored.
2. The features are ranked according to their Fisher ranking score
3. The dataset is randomly shuffled and divided 10 times into the train-test subsets with the ratio 2/1.
4. For each classification algorithm, metrics to maximize, and feature type (PIL, SHARP, SXR), the following algorithm is performed:
 1. Select two features with the highest F-score (or the feature of the highest F-score and SWPC prediction probabilities)
 2. Find the classifier parameters which maximize the mean of the metrics (score) across the train-test data sets
 3. Introduce the feature with the next highest F-score and temporarily add it to the previously-considered features
 4. Find the classifier parameters which maximize the mean of the metrics across the train-test data sets
 5. If the score is higher than previously-obtained plus certain threshold, add the feature permanently. Discard it otherwise. Return to the step c.

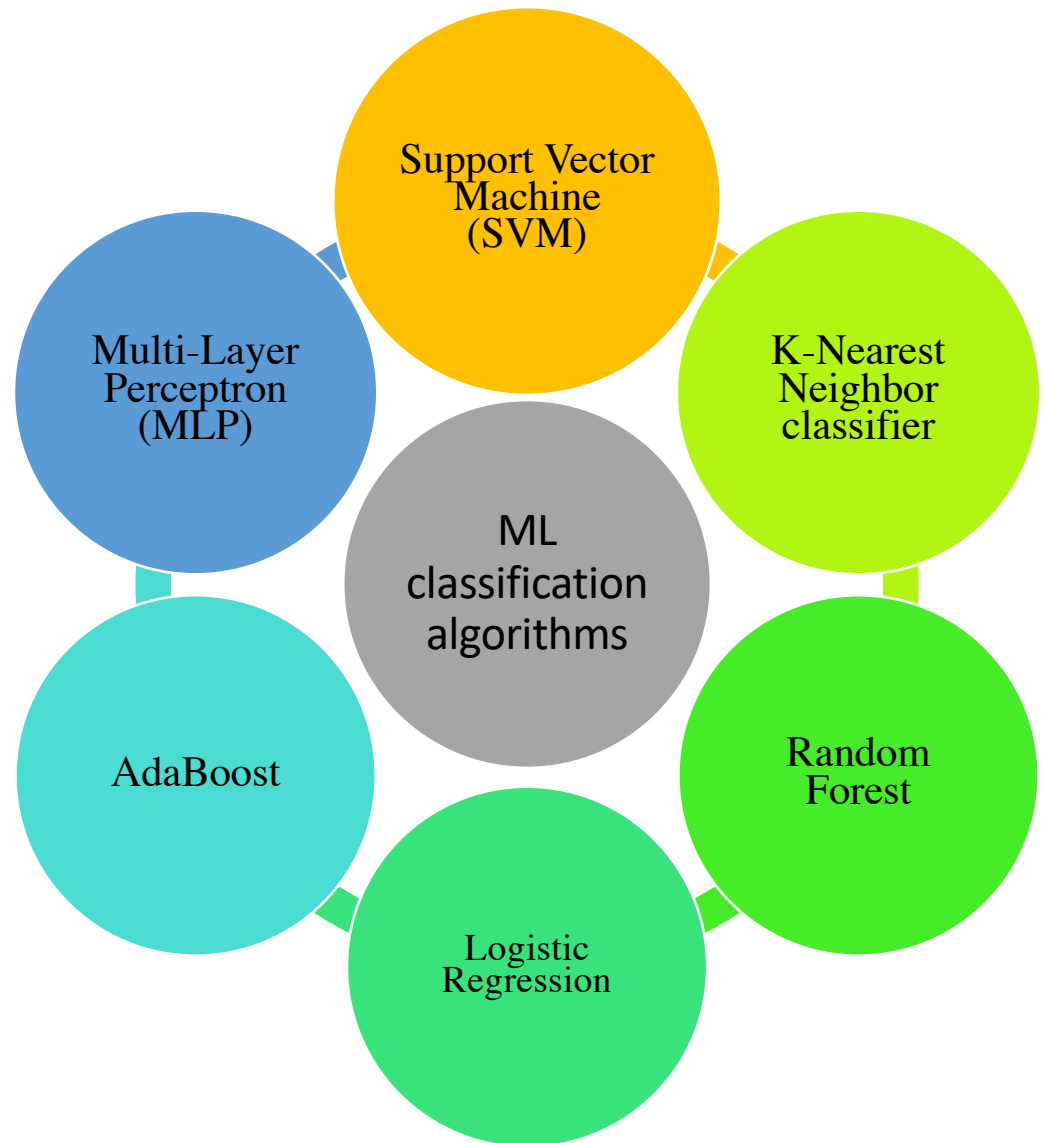
Tested ML algorithms

- Performance of ML algorithms was measured in terms of True Skill Statistics (TSS) and Heidke Skill Score (HSS):

$$\text{TSS} = \frac{\text{TP}}{\text{TP} + \text{FN}} - \frac{\text{FP}}{\text{FP} + \text{TN}}$$

$$\text{HSS} = \frac{2 \times [(\text{TP} \times \text{TN}) - (\text{FN} \times \text{FP})]}{(\text{TP} + \text{FN}) \times (\text{FN} + \text{TN}) + (\text{TP} + \text{FP}) \times (\text{TN} + \text{FP})}$$

- For each algorithm and metrics, we find optimal parameters which maximize the score averaged over train-test subset pairs
- The performance is tested for each group of parameters (PIL/SHARP/SXR/ALL) and including / excluding SWPC probabilities



Enhancement of the Binary (Yes/No) Forecast

We investigate the possibility to enhance the binary (yes/no) forecast of M-class and X-class flares by combining the SWPC NOAA expert predictions (probabilities) with various features (SXR, PIL, SHARP) :

- Support Vector Machine Classifiers (SVMC, SVC) perform better than other considered machine-learning algorithms/classifiers (k-Nearest Neighbor, Random Forest, Neural Networks of different architecture)
- The classifier trained on just one of the feature group (SXR, PIL, SHARP) performs at the same level as expert predictions/probabilities
- The classifier trained on all available features except SWPC NOAA expert predictions significantly outperforms the SWPC NOAA expert predictions in terms of TSS and HSS

TSS, M flares	ES threshold	SVM Linear	SVM RBF	SVM Sigmoid	Logistic Regression	kNN	RF	AdaBoost	NNA1	NNA2	NNA3
Expert scores (ES)	0.560±0.017	-	-	-	-	-	-	-	-	-	-
PIL + ES	-	0.601±0.041	0.598±0.035	0.601±0.040	0.595±0.034	0.541±0.033	0.295±0.045	0.589±0.034	0.572±0.042	0.574±0.041	0.577±0.042
SHARP + ES	-	0.560±0.024	0.586±0.030	0.579±0.029	0.583±0.026	0.515±0.045	0.285±0.033	0.564±0.031	0.551±0.041	0.537±0.035	0.506±0.059
SXR + ES	-	0.568±0.040	0.567±0.038	0.568±0.043	0.567±0.042	0.470±0.029	0.233±0.027	0.559±0.045	0.546±0.028	0.549±0.039	0.545±0.039
ALL + ES	-	0.612±0.039	0.632±0.031	0.617±0.041	0.620±0.034	0.550±0.032	0.294±0.049	0.611±0.035	0.563±0.063	0.522±0.042	0.539±0.039
PIL – ES	-	0.587±0.031	0.588±0.037	0.588±0.023	0.594±0.024	0.521±0.028	0.286±0.038	0.572±0.032	0.546±0.039	0.576±0.031	0.575±0.033
SHARP – ES	-	0.573±0.037	0.583±0.034	0.587±0.032	0.584±0.033	0.510±0.040	0.244±0.033	0.572±0.034	0.557±0.033	0.535±0.045	0.527±0.039
SXR – ES	-	0.564±0.043	0.570±0.039	0.569±0.036	0.570±0.035	0.463±0.035	0.216±0.027	0.567±0.039	0.541±0.039	0.543±0.028	0.535±0.042
ALL – ES	-	0.619±0.030	0.627±0.033	0.635±0.041	0.628±0.025	0.553±0.038	0.289±0.046	0.618±0.028	0.564±0.049	0.499±0.052	0.529±0.057

Enhancement of the Binary (Yes/No) Forecast

We investigate the possibility to enhance the binary (yes/no) forecast of M-class and X-class flares by combining the SWPC NOAA expert predictions (probabilities) with various features (SXR, PIL, SHARP) :

- Support Vector Machine Classifiers (SVMC, SVC) perform better than other considered machine-learning algorithms/classifiers (k-Nearest Neighbor, Random Forest, Neural Networks of different architecture)
- The classifier trained on just one of the feature group (SXR, PIL, SHARP) performs at the same level as expert predictions/probabilities
- The classifier trained on all available features except SWPC NOAA expert predictions significantly outperforms the SWPC NOAA expert predictions in terms of TSS and HSS

HSS, M flares	ES threshold	SVM Linear	SVM RBF	SVM Sigmoid	Logistic Regression	kNN	RF	AdaBoost	NNA1	NNA2	NNA3
Expert scores (ES)	0.412±0.014	-	-	-	-	-	-	-	-	-	-
PIL + ES	-	0.444+0.031	0.445+0.035	0.444+0.029	0.449+0.026	0.396+0.023	0.352+0.044	0.430+0.028	0.428+0.039	0.424+0.028	0.431+0.023
SHARP + ES	-	0.403+0.030	0.426+0.034	0.411+0.042	0.414+0.032	0.372+0.040	0.335+0.028	0.414+0.028	0.400+0.041	0.401+0.051	0.405+0.043
SXR + ES	-	0.417+0.021	0.417+0.019	0.412+0.020	0.426+0.022	0.361+0.035	0.286+0.049	0.410+0.035	0.403+0.024	0.394+0.011	0.386+0.024
ALL + ES	-	0.467+0.040	0.477+0.034	0.467+0.036	0.476+0.031	0.408+0.011	0.350+0.038	0.449+0.024	0.435+0.031	0.441+0.040	0.420+0.053
PIL – ES	-	0.426+0.042	0.430+0.041	0.432+0.041	0.440+0.039	0.377+0.024	0.341+0.047	0.425+0.038	0.401+0.050	0.413+0.033	0.407+0.042
SHARP – ES	-	0.420+0.037	0.439+0.042	0.428+0.042	0.423+0.038	0.362+0.038	0.315+0.030	0.412+0.040	0.386+0.050	0.370+0.040	0.388+0.050
SXR – ES	-	0.406+0.025	0.415+0.027	0.412+0.019	0.416+0.018	0.332+0.028	0.268+0.030	0.398+0.027	0.398+0.019	0.379+0.046	0.396+0.010
ALL – ES	-	0.485+0.038	0.488+0.036	0.482+0.036	0.480+0.030	0.400+0.029	0.364+0.035	0.457+0.038	0.435+0.044	0.431+0.042	0.443+0.042

Enhancement of the Binary (Yes/No) Forecast

We investigate the possibility to enhance the binary (yes/no) forecast of M-class and X-class flares by combining the SWPC NOAA expert predictions (probabilities) with various features (SXR, PIL, SHARP) :

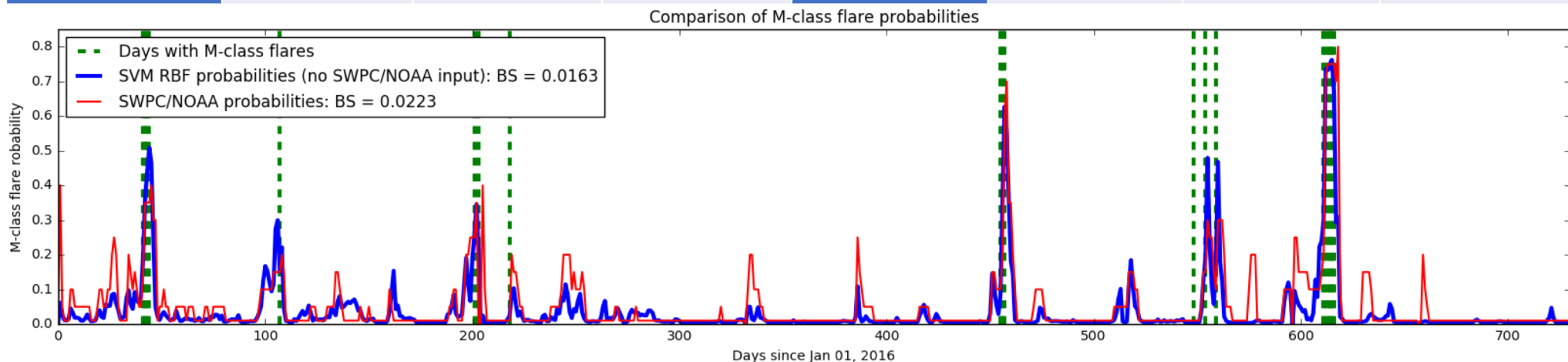
- Support Vector Machine Classifiers (SVMC, SVC) perform better than other considered machine-learning algorithms/classifiers (k-Nearest Neighbor, Random Forest, Neural Networks of different architecture)
- The classifier trained on just one of the feature group (SXR, PIL, SHARP) performs at the same level as expert predictions/probabilities
- The classifier trained on all available features except SWPC NOAA expert predictions significantly outperforms the SWPC NOAA expert predictions in terms of TSS and HSS

TSS, X flares	ES threshold	SVM Linear	SVM RBF	SVM Sigmoid	Logistic Regression	kNN	RF	AdaBoost	NNA1	NNA2	NNA3
Expert scores (ES)	0.575+0.079	-	-	-	-	-	-	-	-	-	-
PIL + ES	-	0.610+0.126	0.605+0.150	0.679+0.129	0.651+0.153	0.583+0.095	0.071+0.075	0.676+0.107	0.352+0.218	0.135+0.141	0.254+0.146
SHARP + ES	-	0.505+0.226	0.556+0.166	0.608+0.179	0.653+0.156	0.367+0.168	0.031+0.067	0.563+0.156	0.164+0.093	0.154+0.137	0.199+0.187
SXR + ES	-	0.722+0.074	0.735+0.070	0.737+0.081	0.753+0.084	0.707+0.211	0.044+0.103	0.780+0.063	0.444+0.211	0.282+0.172	0.221+0.272
ALL + ES	-	0.743+0.073	0.735+0.070	0.768+0.070	0.753+0.084	0.707+0.211	0.044+0.103	0.780+0.063	0.132+0.134	0.111+0.106	0.040+0.072
PIL – ES	-	0.664+0.086	0.659+0.080	0.667+0.039	0.678+0.111	0.411+0.129	0.084+0.091	0.675+0.121	0.368+0.288	0.117+0.188	0.207+0.270
SHARP – ES	-	0.416+0.238	0.413+0.216	0.464+0.165	0.456+0.132	0.410+0.193	0.036+0.073	0.422+0.165	0.129+0.097	0.096+0.132	0.107+0.068
SXR – ES	-	0.774+0.066	0.755+0.078	0.761+0.069	0.757+0.076	0.581+0.146	0.047+0.077	0.780+0.063	0.494+0.205	0.321+0.214	0.406+0.219
ALL – ES	-	0.774+0.066	0.782+0.080	0.761+0.069	0.771+0.073	0.720+0.146	0.051+0.082	0.780+0.063	0.152+0.130	0.039+0.082	0.057+0.087

Enhancement of the Probabilistic Forecast

- Performance of the probabilistic forecast can be measured by Brier Skill Score $BS = \frac{1}{n} \sum_{i=1}^n (P_i - Q_i)^2$
- Probabilities estimated by Support Vector Classifiers (Platt 1999) trained on all features except the SWPC NOAA expert predictions have lower BS (give better prediction) than expert-based probabilities.
- Operational probabilistic prediction for 2016-2017 also has lower BS than the SWPC NOAA predictions

BS, M flares	Expert scores (ES)	SVM RBF	SVM Sigmoid	BS, X flares	Expert Scores (ES)	SVM RBF	SVM Sigmoid
Expert scores (ES)	0.0918+0.0041	-	-	Expert scores (ES)	0.0111+0.0012	-	-
ALL + ES	-	0.0728+0.0043	0.0728+0.0043	ALL + ES	-	0.0067+0.0013	0.0066+0.0013
ALL – ES	-	0.0719+0.0042	0.0720+0.0042	ALL – ES	-	0.0067+0.0013	0.0062+0.0013



The background of the slide is a large, textured blue sphere, resembling a planet or a celestial body, with a white circular area on the right side. The sphere has a mottled, cloudy appearance with various shades of blue and white. The white circle is positioned on the right side of the sphere, and the word "Conclusions" is written in a black serif font within it. A small horizontal line is centered below the word.

Conclusions

IMDSF (<https://heliportal.nas.nasa.gov>) is a fully-functional database of solar flares which:

- Integrates various flare lists and catalogs together with flare-related events
- Identifies uniquely-matched flare events based on time and position information
- Allows to search for the flare events based on their physical descriptors and observational coverage

We are currently working on IDSEAR database which:

- Combines solar events (flares, CMEs, eruptions etc) with properties of solar active regions and observational coverage
- Provides an opportunity for combined queries of these parameters
- (Planned) provides the user with unique AR descriptors and data products

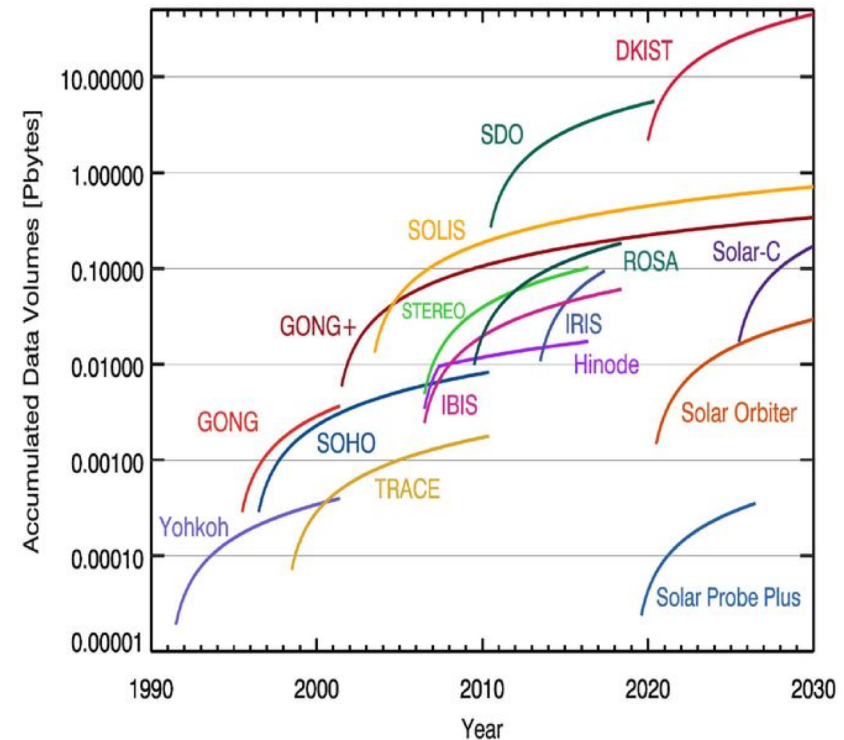
ML forecasts of the solar flares demonstrate promising results:

- Binary (yes/no) forecasts of M/X-class flares based on each feature group (PIL, SHARP, SXR) separately have the same or better performance than the SWPC NOAA operational forecasts.
- It is possible to enhance binary forecast of M/X-class flares by considering joint magnetic (PIL, SHARP) and Soft X-ray characteristics.
- Probabilistic forecast of M-class and X-class flares based on the Support Vector Machine is better than the SWPC NOAA operational forecasts in terms of Brier Skill Score.

A combination of comprehensive data integration and representation techniques and advanced machine learning algorithms is required for accurate prediction of solar activity and data discovery in solar physics

We are living in an Era of large undiscovered scientific data volumes.
We should use this advantage.

Solar Physics Mission Data Sizes



Courtesy K. Reardon



Thank You for
Your Attention!